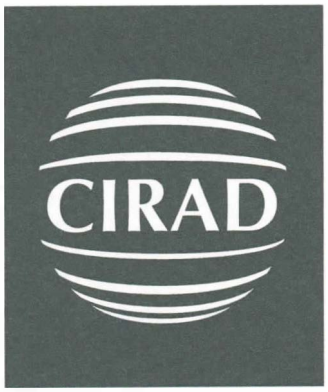


Biométrie

Éléments pour une introduction au Jackknife

J.-C. Bergonzini
H. Ledoux
CIRAD-Forêt

Septembre, 1993

Sommaire

Introduction

Le "Jackknife"[1] ou "Eustachage" en français, est le nom donné par Tukey [11] à la méthode permettant d'estimer la variance d'une statistique par réemploi de l'échantillon. Cette méthode repose sur une technique de réduction du biais des estimateurs paramétriques suggérée par Quenouille [9]. Ce rapport présentera les différentes utilisations de cette méthode :

- La réduction du biais des estimateurs
- L'estimation de la variance de l'estimateur et de son intervalle de confiance
- La détection des valeurs aberrantes

[1] Traduction : Couteau à plusieurs lames, ou couteau Suisse.

Chapitre 1

Estimateur du JACKKNIFE : Réduction du biais.

Soit X une variable aléatoire de fonction de répartition $\mathcal{F}$. Nous pouvons lui associer la suite $(X_1, X_2, \ldots, X_n)$ de n variables aléatoires indépendantes de même loi que X (nous parlerons d'un n-échantillon issu de la variable parente X).

La réalisation de l'échantillon $(X_1, X_2, \ldots, X_n)$ sera notée $(x_1, x_2, \ldots, x_n)$.

Généralement la loi de X sera fonction d'un paramètre θ que nous chercherons à estimer, au moyen d'un estimateur :

$$T(X_1, X_2, \ldots, X_n)$$

la valeur prise par cet estimateur étant l'estimation de θ :

$$\hat{\theta} = T(x_1, x_2, \ldots, x_n)$$

Remarque : Nous parlerons aussi, pour désigner T, de statistique.

Si nous désignons par $E\left[T\left(X_1, X_2, \ldots, X_n\right)\right]$, l'espérance de la statistique T, (lorsqu'elle existe), par rapport à la loi $\mathcal{F}$, la valeur de l'écart :

$$E\left[T\left(X_1, X_2, \ldots, X_n\right)\right] - \theta$$

est le biais de T.

Nous noterons parfois T_n l'estimateur $T(X_1, X_2, \ldots, X_n)$ et
$T_{(-i)} = T(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ l'estimateur calculé sur tous les X_j sauf X_i.

1.1 Le biais de quelques estimateurs

Dans de nombreux cas, nous pouvons montrer que le biais d'un estimateur T_n est de la forme :

$$E(T_n) - \theta = b_n = \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots$$

Les a_i étant indépendants de n.

a l'estimateur de la variance

Soit $V_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ l'estimateur du maximum de vraisemblance de la variance γ.
Nous montrons aisément que :

$$E(V_n) = \gamma - \frac{1}{n}\gamma \quad \Rightarrow \quad b_n = -\frac{1}{n}\gamma$$

D'où : $a_1 = -\gamma \quad a_2 = 0 \cdots$

b l'estimateur de l'écart-type

Nous considérons un échantillon de n variables aléatoires identiques et indépendantes $(X_1, X_2, \ldots, X_n)$ avec :

$$\forall i, \qquad X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

Nous savons que l'estimateur du maximum de vraisemblance de σ^2 s'écrit :

$$\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2$$

qu'il est biaisé et que nous utilisons classiquement :

$$S^2 = \frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2$$

avec :

$$E\left(S^2\right) = \sigma^2 \quad \text{et} \quad \frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$$

Pour estimer σ, nous utilisons généralement la statistique S mais il est clair que $E(S)$ est différent de σ puisque :

$$E\left(\sqrt{S^2}\right) \neq \sqrt{E\left(S^2\right)}$$

Nous pouvons d'ailleurs montrer que :

$$E(S) = \sigma c_n \quad \text{avec} \quad c_n = \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma\left(\frac{1}{2}n\right)}{\Gamma\left(\frac{1}{2}n - \frac{1}{2}\right)}$$

Quelques valeurs de c_n :

n	4	6	8	10	12	16	20	30	50	100
c_n	0,9213	0,9515	0,9650	0,9727	0,9776	0,9835	0,9869	0,9914	0,9949	0,9975

c l'estimateur de la borne d'une loi uniforme

Soit X distribuée de façon uniforme sur $[0, \theta]$ et $(x_1, \ldots, x_n)$ une réalisation de l'échantillon $(X_1, X_2, \ldots, X_n)$. Un estimateur de θ peut être :

$$S_n = \sup_i X_i$$

Nous pouvons montrer alors :

$$E(S_n) - \theta = -\frac{\theta}{n+1} = \frac{\theta}{n} - \frac{\theta}{n^2} + \frac{\theta}{n^3} - \cdots$$

D'où : $a_1 = \theta \quad a_2 = -\theta \cdots$

d l'estimateur du coefficient de corrélation

Soit $R = \dfrac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}}$ l'estimateur du coefficient de corrélation ρ entre deux variables aléatoires normales X, Y.

Si nous notons (Kendall - Stuart [6]) :

$$F(\alpha, \beta, \gamma, t) = 1 + \frac{\alpha\beta}{\gamma}\frac{t}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{t^2}{2!} + \cdots$$

nous pouvons montrer :

$$\begin{aligned} E(R) &= \frac{\rho\Gamma^2\left(\frac{1}{2}n\right)}{\Gamma\left(\frac{1}{2}(n-1)\right)\Gamma\left(\frac{1}{2}(n+1)\right)} F\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}(n+1), \rho^2\right) \\ &= \rho\left(1 - \frac{1-\rho^2}{2n} + o\left(\frac{1}{n^2}\right)\right) \end{aligned}$$

Rappel :

$$\begin{cases} \Gamma(u) = \displaystyle\int_0^\infty t^{u-1}\exp(-t)dt \\ o\left(\dfrac{1}{n^2}\right) \text{ désigne une fonction telle que } \dfrac{o\left(\frac{1}{n^2}\right)}{\frac{1}{n^2}} \to 1 \text{ quand } n \to \infty \end{cases}$$

e l'estimateur d'un quotient

Soit X et Y deux variables aléatoires, auxquelles nous associons les suites $(X_1, X_2, \ldots X_n)$ et $(Y_1, Y_2, \ldots Y_n)$. Nous poserons $\mu = E(X_i)$ et $\eta = E(Y_i)$. Le problème est de vouloir estimer le rapport $\theta = \dfrac{\mu}{\eta}$, car l'estimateur $R = \dfrac{\sum Y_i}{\sum X_i} = \dfrac{\bar{Y}}{\bar{X}}$, est un estimateur biaisé. Nous pouvons montrer que :

$$E(R) = \theta - \frac{\operatorname{cov}(R, \bar{X})}{E(\bar{X})}$$

Certaines études ont été faites pour évaluer $E(R)$ dans différents cas. Par exemple Durbin [2] étudie la forme de $E(R)$ lorsque la régression entre X et Y est linéaire et X distribuée de façon normale. Il montre que, dans un tel cas :

$$E(R) - \theta = a\left(\frac{1}{n} + \frac{3}{n^2} + \frac{15}{n^3}\right) + o\left(\frac{1}{n^4}\right)$$

De nombreuses personnes ont proposé d'autres estimateurs pour estimer un rapport [7], et même si le Jackknife n'est pas toujours le meilleur estimateur, il s'en approche beaucoup.

f remarque

Il est bien évident que tous les estimateurs n'ont pas un biais qui peut s'exprimer sous cette forme.

Le problème de la structure de T ne semble pas complètement éclairci.

1.2 L'estimateur du JACKKNIFE

L'idée de base du Jackknife provient du fait que la donnée des $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ donc d'un n-échantillon, est la donnée de plusieurs $(n-r)$-échantillons.

Intuitivement, cela devrait apporter une information sur les biais : $b_1, b_2, \ldots, b_n$.

Il y a donc peut-être moyen de calculer ainsi, certains des paramètres $a_1, a_2, \ldots$ du développement :

$$b_n = \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots$$

Les a_i étant indépendants de n.

Notations :

$$\begin{array}{ll} T & \text{l'estimateur } T_n(X_1, X_2, \ldots\ldots X_n) \\ T_{(-i)} & \text{l'estimateur } T_{n-1}(X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots X_n) \\ T_{(.)} & = \frac{1}{n}\sum_{i=1}^{n} T_{(-i)} \end{array}$$

$$\begin{array}{ll} \hat{\theta} & \text{la valeur prise par } T_n(x_1, x_2, \ldots x_n) \\ \hat{\theta}_{(-i)} & \text{la valeur prise par } T_{n-1}(x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots x_n) \end{array}$$

Considérons un estimateur de θ à partir d'un $(n-1)$-échantillon :

$$E\left[T(X_1, X_2, \cdots X_{i-1}, X_{i+1}, \cdots X_n)\right] = \theta + b_{n-1}$$

et nous pouvons construire n estimateurs de ce type, en retranchant X_i à l'échantillon $(X_1, X_2, \ldots X_n)$.

Nous avons : $\begin{cases} E(T) & = \theta + \dfrac{a_1}{n} + \dfrac{a_2}{n^2} + \cdots \\ E\left(T_{(.)}\right) & = \theta + \dfrac{a_1}{n-1} + \dfrac{a_2}{(n-1)^2} + \cdots \end{cases}$

L'estimateur du Jackknife $\tilde{T}$ est une combinaison linéaire de T et $T_{(.)}$.

$$\boxed{\tilde{T} = nT - (n-1)T_{(.)}}$$

$$\begin{aligned} E(\tilde{T}) &= n\theta + a_1 + \frac{a_2}{n} + \frac{a_3}{n^2} + \cdots - \left((n-1)\theta + a_1 + \frac{a_2}{n-1} + \cdots\right) \\ &= \theta + a_2\left(\frac{1}{n} - \frac{1}{n-1}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \cdots \\ &= \theta - \frac{a_2}{n(n-1)} + \frac{1-2n}{n^2(n-1)^2}a_3 + \cdots \\ &= \theta + o\left(\frac{1}{n^2}\right) \end{aligned}$$

Conclusion : Asymptotiquement, le biais de $\tilde{T}$ est moindre que celui de T .

Remarque :

$$\begin{aligned}\tilde{T} &= nT - (n-1)T_{(.)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[T - (n-1)\left(T_{(-i)} - T\right)\right]\end{aligned}$$

Les valeurs $T - (n-1)\left(T_{(-i)} - T\right)$ lorsque i varie de 1 à n, sont les **"pseudo-valeurs"** introduites par Tukey [11]. Elles sont notées $\tilde{T}_{(i)}$.

$$\boxed{\tilde{T} = \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_{(i)}}$$

Nous noterons les valeurs prises par ces "pseudo-valeurs" :

$$\tilde{\theta}_{(i)} = \hat{\theta} - (n-1)\left(\hat{\theta}_{(-i)} - \hat{\theta}\right) = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

et $\tilde{\theta}$ la valeur prise par $\tilde{T}$ l'estimateur du Jackknife qui est la moyenne de ces "pseudo-valeurs".

1.3 Des exemples d'estimateurs

a l'estimateur de la variance

Soit la réalisation d'un échantillon :

i	1	2	3	4	5
x_i	$-1,0$	$0,0$	$0,5$	$1,0$	$4,0$

la valeur prise par l'estimateur de la variance est :

$$\hat{\gamma} = \frac{1}{n}\sum(x_i - \bar{x})^2 = 2,84$$

les valeurs prises par cet estimateur sur un 4-échantillon :

$\hat{\gamma}_{(-i)}$	$2,42$	$3,29$	$3,50$	$3,55$	$0,54$

Calculons les "pseudo-valeurs" : $\tilde{\gamma}_{(i)} = 5\hat{\gamma} - 4\hat{\gamma}_{(-i)}$

$\tilde{\gamma}_{(i)}$	$4,52$	$1,04$	$0,20$	$0,00$	$12,04$

La moyenne de ces "pseudo-valeurs" est l'estimateur du Jackknife :

$$\tilde{\gamma} = \frac{1}{5}\sum_{i=1}^{5}\tilde{\gamma}_{(i)} = 3,56$$

b l'estimateur de l'écart-type

Pour étudier l'estimateur du Jackknife, nous allons procéder par simulation des variables :

$$\forall i, \quad X_i \sim \mathcal{N}(0,10) \quad \text{et} \quad \sigma = \sqrt{10} = 3,1623$$

Soit un échantillon $(x_1, x_2, \ldots, x_n)$, nous calculons :

$$\hat{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 \quad , \quad \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \quad \text{et} \quad \hat{s}$$

$$\hat{s}^2_{(-i)} = \frac{1}{n-2}\sum_{j\neq i}\left(x_j - \bar{x}_{(-i)}\right)^2 \quad , \quad \bar{x}_{(-i)} = \frac{1}{n-1}\sum_{j\neq i} x_j \quad \text{et} \quad \hat{s}_{(-i)}$$

les "pseudo-valeurs" : $\tilde{s}_{(i)} = n\hat{s} - (n-1)\hat{s}_{(-i)}$

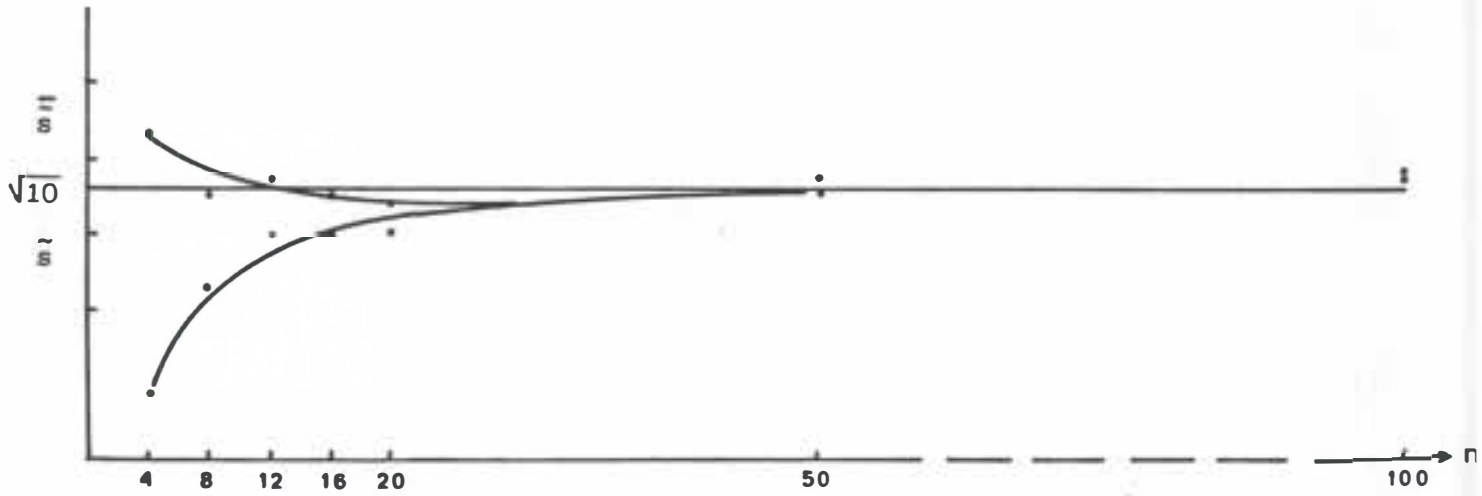
et l'estimateur du Jackknife : $\tilde{s} = \frac{1}{n}\sum_{i=1}^{n}\tilde{s}_{(i)}$

Nous répétons l'opération 500 fois et nous obtenons les distributions empiriques et simulées de $\hat{s}$ et $\tilde{s}$:

$$\begin{array}{cccccc} \hat{s}_1 & \hat{s}_2 & \hat{s}_3 & \ldots & \hat{s}_{500} \\ \tilde{s}_1 & \tilde{s}_2 & \tilde{s}_3 & \ldots & \tilde{s}_{500} \end{array}$$

Moyenne des $\hat{s}_i$ et des $\tilde{s}_i$ pour différentes valeurs de n ($\sigma = 3,1623$) :

n	4	8	12	16	20	50	100
$\bar{\hat{s}}$	2,8896	3,0315	3,0994	3,0990	3,0984	3,1505	3,1717
$c_n\sigma$	2,9134	3,0516	3,0914	3,1101	3,1209	3,1462	3,1543
$\bar{\tilde{s}}$	3,2283	3,1541	3,1744	3,1524	3,1414	3,1667	3,1798
$\lvert\bar{\hat{s}} - \sigma\rvert$	0,2727	0,1107	0,0633	0,0633	0,0639	0,0118	0,0094
$\lvert\bar{\tilde{s}} - \sigma\rvert$	0,0660	0,0082	0,0121	0,0099	0,0209	0,0044	0,0175

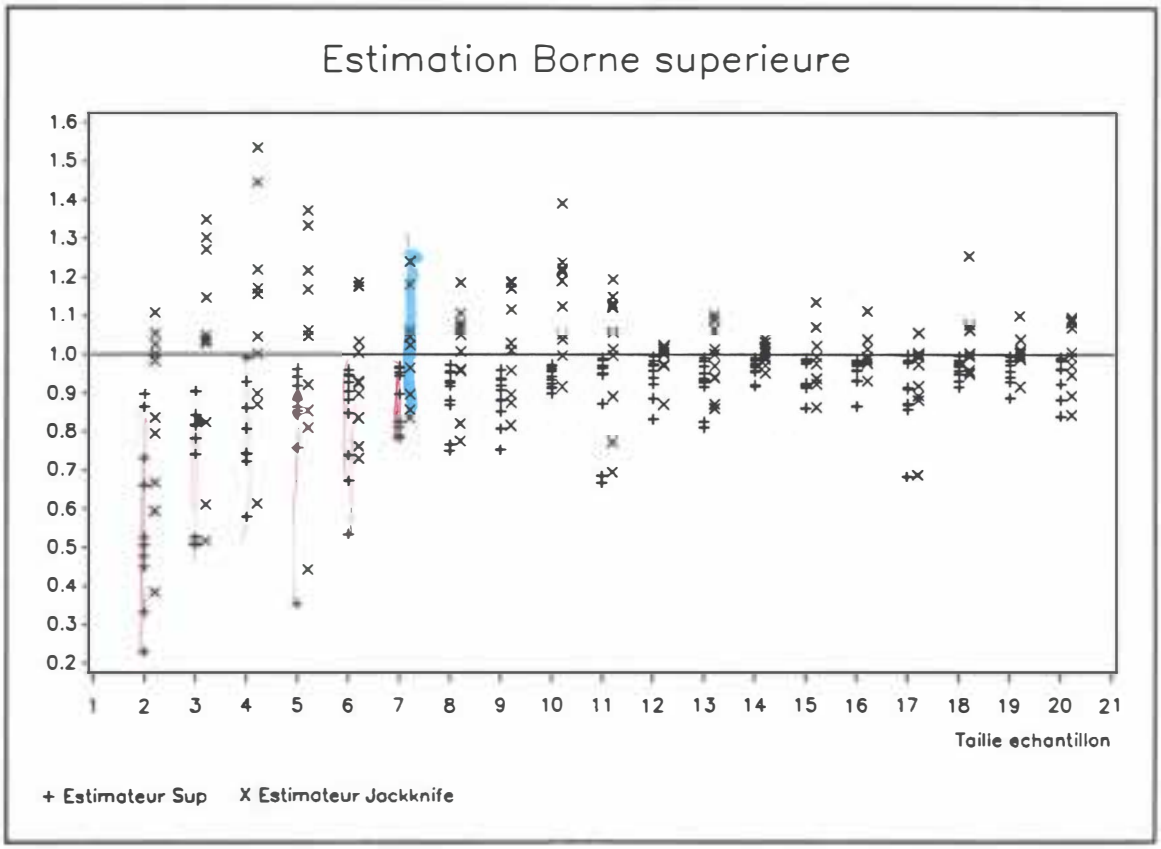


c l'estimateur de la borne d'une loi uniforme

Soit X une variable aléatoire distribuée de façon uniforme sur $[0, \theta]$.

Dans notre exemple, nous prendrons $\theta = 1$. Nous tirerons aléatoirement une dizaine d'échantillons de taille n et nous ferons varier n de 2 à 20.
Pour chacun de ces échantillons, nous calculerons la valeur prise par l'estimateur $\hat{\theta} = \sup_i x_i$, et par l'estimateur du Jackknife $\tilde{\theta}$.

Les valeurs obtenues sont présentées sur un graphe dont les abscisses correspondent à la taille de l'échantillon.



Nous remarquons :

- Les valeurs prises par l'estimateur sont toujours plus faibles que la valeur de la borne. La différence, ou le biais, diminuant lorsque n augmente.

- Les valeurs prises par l'estimateur du Jackknife semblent moins biaisées que l'estimateur du *sup*, mais celles-ci sont, par contre, plus étalées.

d l'estimateur du coefficient de corrélation

Etudions la relation entre le taux de goudron (X) et le taux de nicotine (Y) [1]. L'échantillon mesuré porte sur 10 marques de cigarettes :

[1] Selon la loi n° 91.32 "Nuit gravement à la santé"

x_i	0,45	0,77	1,07	1,03	1,34	1,14	1,15	0,90	0,55	1,15
y_i	11,0	13,0	14,0	15,0	17,0	18,0	14,5	13,5	8,5	16,5

La valeur prise par l'estimateur du coefficient de corrélation est $\hat{\rho} = 0,89$.

$\hat{\rho}_{(-i)}$	0,90	0,89	0,90	0,89	0,88	0,91	0,91	0,89	0,87	0,88
$\tilde{\rho}_{(i)}$	0,79	0,89	0,79	0,89	0,98	0,71	0,71	0,89	1,07	0,98

La valeur prise par l'estimateur du Jackknife est $\tilde{\rho} = \frac{1}{n}\sum \tilde{\rho}_{(i)} = 0,87$

e l'estimateur d'un quotient

Etudions le taux de surface boisée en France. Les données recueillies proviennent de l'inventaire forestier national de 1983 [1], et correspondent aux données de 90 départements[2]. Nous connaissons la surface de la France, qui est de 54.918.450 ha[3]. Nous connaissons aussi exactement la superficie totale boisée, qui est de 13.775.471 ha, d'où le taux de boisement national de 25,0835 %.

Oublions les données précédentes. Nous connaissons toujours la surface de la France, mais nous voulons estimer la superficie nationale boisée ou, ce qui revient au même, le taux de boisement national. Pour cela nous utiliserons l'estimateur du quotient $Q = \frac{\sum Y_i}{\sum X_i} = \frac{\bar{Y}}{\bar{X}}$. Notre échantillon sera composé de quelques départements français.

Soit 10 départements tirés au hasard :

Code	Département	Surface boisée	Superficie du département	Taux de boisement
01	Ain	177.071	578.501	30,61
12	Aveyron	224.603	877.122	25,61
17	Charente Maritime	101.706	690.000	14,74
34	Hérault	138.484	622.673	22,24
37	Indre et Loire	135.776	615.403	22,06
44	Loire Atlantique	42.515	695.640	6,11
54	Meurthe et Moselle	169.201	527.737	32,06
74	Haute – Savoie	170.814	483.862	35,30
83	Var	280.016	603.250	46,42
87	Haute – Vienne	135.210	555.825	24,33
Total		1.575.396	6.250.013	

Calculons la valeur prise par l'estimateur du quotient :

$$\hat{q} = \frac{1.575.396}{6.250.013} * 100\ \% = 25,206\ \%$$

[2]la Corse et l'Île-de-France sont comptées chacune pour deux départements

[3]Cette superficie diffère de la superficie nationale officielle de 54.919.193 ha, car les données de surface de chaque département sont celles des différentes années d'inventaire

Calculons cet estimateur en ôtant un par un les départements, et calculons les "pseudo-valeurs" :

Code	$\tilde{q}_{(-i)}$	$\tilde{q}_{(i)}$
01	24,655	30,166
12	25,141	25,794
17	26,505	13,516
34	25,534	22,252
37	25,550	22,116
44	27,598	3,683
54	24,574	30,896
74	24,359	32,831
83	22,940	45,601
87	25,292	24,433
Moyenne		25,129

La valeur prise par l'estimateur du Jackknife est :

$$\tilde{q} = \frac{1}{10}\sum_{i=1}^{10} \tilde{q}_{(i)} = 25,129\ \%$$

Ce résultat est à comparer avec un autre estimateur $R = \frac{1}{n}\sum \frac{Y_i}{X_i}$, qui représente le taux moyen de boisement par département.
La valeur prise par cet estimateur est : $\hat{r} = 25,948\ \%$.

1.4 Le Jackknife et le biais

a le Jackknife et les estimateurs sans biais

Supposons que l'estimateur du Jackknife $\tilde{T}$ prenne les mêmes valeurs que l'estimateur T $\left(T \equiv \tilde{T}\right)$, nous avons :

$$\begin{aligned} T = \tilde{T} &= nT - (n-1)T_{(.)} \\ T &= T_{(.)} \\ E(T) &= E\left(T_{(.)}\right) \end{aligned}$$

Nous savons, dans ce cas, que l'espérance d'un estimateur construit sur un n-échantillon est égale à l'espérance d'un estimateur construit sur un (n-1)-échantillon.

$$E(T_n) = E\left(T_{(.)}\right) = E(T_{n-1})$$

Conclusion :

Si $\tilde{T} \equiv T$ quel que soit n, $E(T_n) = E(T_m)$ quels que soient n et m.
si l'estimateur est convergent : il est sans biais.
Si $E(T) = \theta$, il est évident que $E\left(\tilde{T}\right) = \theta$.

b estimation du biais

Quenouille [10] propose d'utiliser $(n-1)\left(T_{(.)}-T\right)$ comme estimateur du biais de T.

$$\tilde{T} = nT - (n-1)T_{(.)} = T - \underbrace{(n-1)\left(T_{(.)}-T\right)}_{\text{correction apportée à } T}$$

Nous avons :

$$E\left(T_{(.)}-T\right) = \underbrace{\frac{a_1}{n-1}+\frac{a_2}{(n-1)^2}+\cdots}_{\text{biais } b_{n-1}} - \underbrace{\frac{a_1}{n}+\frac{a_2}{n^2}+\cdots}_{\text{biais } b_n}$$

Si nous notons B_n l'estimateur du biais b_n :

$$\begin{aligned} E(B_n) &= (n-1)(b_{n-1}-b_n) \\ E(B_n)-b_n &= (n-1)b_{n-1}-nb_n \end{aligned}$$

Pour que ce biais soit nul, il faut que $b_n = \frac{n-1}{n}b_{n-1}$ et pour que ceci soit vérifié, pour tout n il faut que b_n soit fonction du type $\frac{a_1}{n}$.

c ce n'est pas l'estimateur du moindre biais

L'estimateur du Jackknife est une combinaison linéaire de T et $T_{(.)}$.

$$\tilde{T} = nT - (n-1)T_{(.)}$$

mais "a priori" ce n'est pas l'estimateur du moindre biais que l'on peut construire ainsi.

$$\begin{aligned} E(T) &= \theta + \frac{a_1}{n}+\frac{a_2}{n^2}+\frac{a_3}{n^3}+\cdots \\ E\left(T_{(.)}\right) &= \theta + \frac{a_1}{n-1}+\frac{a_2}{(n-1)^2}+\frac{a_3}{(n-1)^3}+\cdots \end{aligned}$$

Remarque :

$$\begin{aligned} \frac{1}{n-1} &= \frac{1/n}{1-1/n} &&= \frac{1}{n}\left(1+\frac{1}{n}+\frac{1}{n^2}+\frac{1}{n^3}+\cdots\right) \\ \frac{1}{(n-1)^2} &= \frac{1/n^2}{(1-1/n)^2} &&= \frac{1}{n^2}\left(1+\frac{2}{n}+\frac{3}{n^2}+\frac{4}{n^3}+\cdots\right) \\ \frac{1}{(n-1)^3} &= \frac{1/n^3}{(1-1/n)^3} &&= \frac{1}{n^3}\left(1+\frac{3}{n}+\frac{6}{n^2}+\frac{10}{n^3}+\cdots\right) \end{aligned}$$

ce qui permet de réécrire $E\left(T_{(.)}\right)$:

$$E\left(T_{(.)}\right) = \theta + \frac{a_1}{n}+\frac{a_1+a_2}{n^2}+\frac{a_1+2a_2+a_3}{n^3}+\cdots$$

Considérons $\alpha T + \beta T_{(.)}$; nous voyons qu'il est nécessaire de prendre $\alpha+\beta=1$ pour que l'estimateur $\alpha T+\beta T_{(.)}$ converge vers θ lorsque $n \to \infty$.

Prenons $\alpha+\beta=1$ et $\alpha = un+v$ (linéaire en n).

$$E\left(\alpha T+\beta T_{(.)}\right) = \theta + \frac{1}{n}a_1(1-u)+\frac{1}{n^2}(-va_1+a_2)+\cdots$$

Il suffit de prendre $u = 1$ pour éliminer les termes en $\frac{1}{n}$. Si le rapport $\frac{a_1}{a_2}$ était connu $\alpha = n - \frac{a_2}{a_1}$ permettrait d'éliminer aussi les termes en $\frac{1}{n^2}$. Dans le cas du Jackknife, on prend $v = 0$.

1.5 Variantes sur le Jackknife

a élimination des termes en $1/n^2$

Essayons de construire un estimateur dont l'espérance soit de la forme : $\theta + o\left(\frac{1}{n^3}\right)$.
Ce problème a été résolu par Gray et Schucany [4]. L'idée est la suivante : pour éliminer les termes en $\frac{1}{n}$ il faut une combinaison de T et $T_{(.)}$, pour éliminer les termes en $\frac{1}{n}$ et $\frac{1}{n^2}$, nous utiliserons une combinaison entre $T, T_{(.)}$ et $T_{(..)}$ où :

$$T_{(..)} = \frac{1}{C_n^2} \sum_{i,j} T_{(-ij)}$$

$T_{(-ij)}$ étant l'estimateur obtenu en éliminant X_i et X_j du n-échantillon $(X_1, X_2, \ldots X_n)$.
Autrement dit :

$$E\left(T_{(-ij)}\right) = E(T..) = \theta + \frac{a_1}{n-2} + \frac{a_2}{(n-2)^2} + \cdots$$

L'estimateur cherché $\tilde{\tilde{T}}$ s'écrit :

$$\tilde{\tilde{T}} = \begin{vmatrix} T & T_{(.)} & T_{(..)} \\ 1 & 1 & 1 \\ \overline{n} & \overline{n-1} & \overline{n-2} \\ \dfrac{1}{n^2} & \dfrac{1}{(n-1)^2} & \dfrac{1}{(n-2)^2} \\ 1 & 1 & 1 \\ \overline{n} & \overline{n-1} & \overline{n-2} \\ \dfrac{1}{n^2} & \dfrac{1}{(n-1)^2} & \dfrac{1}{(n-2)^2} \end{vmatrix}$$

Après calcul des deux déterminants, nous obtenons :

$$\tilde{\tilde{T}} = \frac{n^2 T - 2(n-1)^2 T_{(.)} + (n-2)^2 T_{(..)}}{2}$$

Nous vérifions que : $E(\tilde{\tilde{T}}) = \theta + o\left(\frac{1}{n^3}\right)$.

Remarque :
Nous pourrions évidemment éliminer les termes en $\frac{1}{n^3}$, $\frac{1}{n^4}, \ldots$ Cette démarche peut paraître alléchante, mais il faut tenir compte :

- des coûts de calcul qui augmentent,
- de l'évolution de la précision des estimateurs.

b le Jackknife de groupe

Admettons que n soit égal au produit de g par h (g et h entiers), nous diviserons le n-échantillon en g sous-ensembles de h-échantillons.
Deux possibilités nous sont offertes pour calculer un estimateur du type Jackknife.

1)

Nous pouvons considérer que les g sous-ensembles déterminent une partition du n-échantillon,

$$\underbrace{x_1\ x_2 \ldots x_h}_{g_1} \quad \underbrace{x_{h+1} \ldots x_{h+h}}_{g_2} \quad \ldots$$

et définir un estimateur du type Jackknife :

$$\check{T} = gT - (g-1)\check{T}_{(.)}$$

avec : $\check{T}_{(.)} = \frac{1}{g}\sum_{i=1}^{g} \check{T}_{(-i)}$

et $\check{T}_{(-i)}$ l'estimateur construit en éliminant les observations du groupe i

Calcul de $E\left(\check{T}_{(.)}\right)$:

$$E\left(\check{T}_{(-i)}\right) = E\left(\check{T}_{(.)}\right) = \theta + \frac{a_1}{n-h} + \frac{a_2}{(n-h)^2} + \ldots$$

$$E\left(\check{T}_{(.)}\right) = \theta + \frac{a_1}{h(g-1)} + \frac{a_2}{(h(g-1))^2} + \ldots$$

$$E(T) = \theta + \frac{a_1}{hg} + \frac{a_2}{(hg)^2} + \ldots$$

$$E(\check{T}) = \theta + g\left(\frac{a_2}{(hg)^2} + \frac{a_3}{(hg)^3} + \ldots\right) - (g-1)\left(\frac{a_2}{(h(g-1))^2} + \frac{a_3}{(h(g-1))^3} + \ldots\right)$$

Les termes en a_1 disparaissent :

$$E(\check{T}) = \theta + a_2\left(\frac{g}{n^2} - \frac{g-1}{(n-1)^2}\right) + a_3\left(\frac{g}{n^3} - \frac{g-1}{(n-h)^3}\right) + \ldots$$

Etudions le terme en a_2 :

$$\frac{g}{n^2} - \frac{g-1}{(n-1)^2} = \frac{g}{n^2} - (g-1)\left(\frac{1}{n^2} + \frac{2h}{n^3} + \frac{3h^2}{n^4} + \ldots\right)$$
$$= -\frac{1}{n^2} - \frac{h}{n^3} + \ldots$$

Nous obtenons un estimateur $\check{T}$ tel que $E(\check{T}) = \theta + o\left(\frac{1}{n^2}\right)$

Quenouille [9] fut le premier à utiliser une telle méthode dans une étude de séries temporelles en prenant le cas où $g = 2$.

2)

Il semble plus logique, bien que plus coûteux en temps de calcul, de faire intervenir tous les estimateurs construits sur $n-h$ observations. Ces estimateurs seront notés $\check{T}_{[-h]}$. Leur moyenne sera égale à : $\check{T}_{[.]} = \frac{1}{C_n^h} \sum_h \check{T}_{[-h]}$ Nous construisons ensuite :

$$\check{T} = gT - (g-1)\check{T}_{[.]}$$

Nous pouvons montrer :

$$E(\check{T}) = \theta + o\left(\frac{1}{n^2}\right)$$

Conclusion :

A ce niveau (celui du biais), il ne semble pas qu'il y ait intérêt à utiliser des Jackknife de groupe, sauf (peut-être) en ce qui concerne le repérage de sous-ensembles de valeurs aberrantes ou bien si l'on désire faire du "Jackknife" à partir d'un grand échantillon.

Chapitre 2

Variance de l'estimateur du JACKKNIFE

2.1 Variance de l'estimateur du Jackknife

Tukey a introduit la notion de "pseudo-valeurs" :

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

l'estimateur du Jackknife apparaissant comme la moyenne de ces "pseudo-valeurs" :

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_{(i)}$$

Il a supputé, semble-t-il par analogie avec la variance d'une moyenne, que :

$$\boxed{s^2 = \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2}$$

était une estimation de la variance de T, et de $\tilde{T}$ l'estimateur du Jackknife.

Par ailleurs, il a fait l'hypothèse que :

$$\frac{\tilde{T}-\theta}{S} \text{ avec } S^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{T}_{(i)} - \tilde{T}\right)^2$$

suivait une loi de Student à (n-1) ddl.

Tukey part de l'idée que les "pseudo-valeurs" $\tilde{T}_{(i)}$ (variables aléatoires qui prennent pour valeur $\tilde{\theta}_{(i)}$) sont indépendantes, ou faiblement corrélées.

Démonstration :

Si nous calculons les covariances entre "pseudo-valeurs", nous obtenons :

$$\begin{aligned}\text{cov}\left(\tilde{T}_{(i)}, \tilde{T}_{(i')}\right) &= \text{cov}\left(nT - (n-1)T_{(-i)}, nT - (n-1)T_{(-i')}\right) \\ &= n^2\text{var}(T) - 2n(n-1)\text{cov}\left(T, T_{(-i)}\right) + (n-1)^2\text{cov}\left(T_{(-i)}, T_{(-i')}\right)\end{aligned}$$

Nous admettrons que les observations sont indépendantes et que la statistique est symétrique : $\text{cov}\left(T, T_{(-i)}\right) = \text{cov}\left(T, T_{(-i')}\right)$.

Considérons une statistique de la forme :

$$T(X_1, X_2, \ldots X_n) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Si nous posons $\gamma = \text{var}(f(X))$, nous avons :

$$\text{var}(T) = \frac{1}{n}\gamma \quad \text{cov}\left(T, T_{(-i)}\right) = \frac{1}{n}\gamma \quad \text{cov}\left(T_{(-i)}, T_{(-i')}\right) = \frac{n-2}{(n-1)^2}\gamma$$

et nous trouvons :

$$\text{cov}\left(\tilde{T}_{(i)}, \tilde{T}_{(i')}\right) = 0$$

Nous obtenons aussi :

$$\text{var}\left(\tilde{T}_{(i)}\right) = \gamma$$

En effet :

$$\begin{aligned}
\tilde{T}_{(i)} &= nT - (n-1)T_{(-i)} \\
\text{var}\left(\tilde{T}_{(i)}\right) &= n^2\text{var}(T) + (n-1)^2\text{var}\left(T_{(-i)}\right) - 2n(n-1)\text{cov}\left(T, T_{(-i)}\right) \\
&= n^2\frac{\gamma}{n} + (n-1)^2\frac{\gamma}{n-1} - 2n(n-1)\frac{\gamma}{n} \\
&= \gamma
\end{aligned}$$

Nous pouvons donc prendre $\frac{1}{n}\sum\left(\tilde{T}_{(i)} - \tilde{T}\right)^2$ comme estimateur de γ

et $\frac{1}{n(n-1)}\sum\left(\tilde{T}_{(i)} - \tilde{T}\right)^2$ comme estimateur de $\text{var}\left(\tilde{T}\right)$

Conclusion :

Il est évident que les statistiques de ce type sont peu nombreuses $\left(\bar{X} = \frac{1}{n}\sum X_i\right)$, par contre, de nombreuses statistiques peuvent être approchées par des estimateurs de cette forme pour n suffisamment grand.

Remarque 1 :

$$s^2 = \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 = \frac{(n-1)}{n}\sum\left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2$$

En effet :

$$\begin{aligned}
\tilde{\theta}_{(i)} &= n - (n-1)\hat{\theta}_{(-i)} \\
\tilde{\theta} &= n\hat{\theta} - \frac{n-1}{n}\sum\hat{\theta}_{(-i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(.)}
\end{aligned}$$

D'où :

$$\tilde{\theta}_{(i)} - \tilde{\theta} = (n-1)\left(\hat{\theta}_{(.)} - \hat{\theta}_{(-i)}\right)$$

Remarque 2 :

Un autre estimateur de la variance a été proposé :

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \hat{\theta}\right)^2$$

Etudions le lien entre s'^2 et s^2 :

$$\begin{aligned}
s'^2 &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \hat{\theta}\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \tilde{\theta} + \tilde{\theta} - \hat{\theta}\right)^2 \\
&= \frac{1}{n(n-1)}\sum\left\{\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 + \left(\tilde{\theta} - \hat{\theta}\right)^2 + 2\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)\left(\tilde{\theta} - \hat{\theta}\right)\right\} \\
&= \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 + \frac{n}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)^2 + \frac{2}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right) \\
&= \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 + \frac{n}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)^2 + \frac{2n}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)\left(\tilde{\theta} - \tilde{\theta}\right)
\end{aligned}$$

D'où :

$$\boxed{s'^2 = s^2 + \frac{1}{n-1}\left(\tilde{\theta} - \hat{\theta}\right)^2}$$

Cet estimateur est considéré comme plus "conservateur" car le terme $\left(\tilde{\theta} - \hat{\theta}\right)^2$ est toujours positif, et donc :

$$s'^2 \geq s^2$$

Remarque 3 :

L'hypothèse avancée par Tukey permet de définir un intervalle de confiance, au niveau $(1-\alpha)$ du paramètre θ :

$$\theta \in \left[\tilde{\theta} - t_{\alpha/2;(n-1)}s \;;\; \tilde{\theta} + t_{\alpha/2;(n-1)}s\right]$$

Remarque 4 :

La conjecture de Tukey a été infirmée par des contre-exemples, dont certains seront présentés dans les paragraphes suivants.
Bien qu'il y ait une certaine robustesse de la loi de Student vis-à-vis de la non-indépendance des échantillons, il faut faire attention au cas de "pseudo-valeurs" discrètes, en particulier dans le cas d'estimation de la médiane ou d'autres statistiques d'ordre.

Une règle pour réduire les degrès de liberté de T s'avère nécessaire : celle proposée par Mosteller et Tukey [8] est simple et utile :

- Compter le nombre de "pseudo-valeurs" réellement différentes ;
- en soustraire un ;
- utiliser le résultat comme degrés de liberté.

2.2 Démarche suivie par Efron

Nous avons vu :

$$s^2 = \frac{1}{n(n-1)} \sum \left(\hat{\theta}_{(i)} - \bar{\theta}\right)^2 = \frac{(n-1)}{n} \sum \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2$$

C'est sous la deuxième forme que l'estimation de la variance est utilisée par Efron [3], il décompose la construction de s^2 en deux étapes :

1. $\sum \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2$ qui est construit en n'utilisant que des $(n-1)$ -échantillons, est considéré comme une estimation de γ_{n-1} la variance de T_{n-1} .
2. $\frac{n-1}{n}$ est alors considéré comme un coefficient correcteur qui permet de passer à l'estimation de γ_n.

Efron démontre, par ailleurs :

$$E\left(\sum_{i=1}^{n} \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_{n-1}$$

La démonstration est développée en annexe.

Remarque 1 :
Efron ne s'intéresse qu'aux statistiques qui sont des fonctions symétriques des X_i.

Remarque 2 :
Si T est de la forme : $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$, les termes en $\sigma_B^2, \ldots$ sont nuls et $\sum_{i=1}^{n} \left(T_{(i)} - T_{(.)}\right)^2$ est un estimateur sans biais de γ_{n-1}.

Remarque 3 :
Généralement, nous n'avons pas :

$$E\left(\frac{n-1}{n} \sum_{i=1}^{n} \left(T_{(i)} - T_{(.)}\right)^2\right) \geq \gamma_n$$

Exemple :

Une U-statistique est une statistique de la forme :

$$U_n(X_1, X_2, \ldots X_n) = \frac{1}{C_n^m} \sum_{C} g(X_{i1}, X_{i2}, \ldots X_{im})$$

où C indique que la sommation s'effectue sur l'ensemble des combinaisons de m entiers choisis parmi $(1, 2, \ldots n)$, de plus g est symétrique et $E(g), E(g^2), \ldots$ existent.

L'estimateur de la variance, par exemple, est une U-statistique :

$$\begin{aligned}\frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2 &= \frac{1}{2(n-1)n}\sum_{i,i'}(X_i - X_{i'})^2 \\ &= \frac{1}{C_n^2}\sum_{i,i'} 4\,(X_i - X_{i'})^2\end{aligned}$$

Hoeffding [5] a montré que pour une U-statistique avec $m \leq n-1$, nous avons :

$$\frac{n-1}{n}\text{var}\,(T_{n-1}) \geq \text{var}\,(T_n)$$

ce qui entraîne :

$$E\left(\frac{n-1}{n}\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_n$$

2.3 Exemples de variances d'estimateurs

a l'estimateur de la moyenne

Soit $\hat{\theta}$, la valeur prise par l'estimateur de la moyenne. Nous savons que :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Nous pouvons montrer aisément que les "pseudo-valeurs" et la valeur prise par l'estimateur du Jackknife s'écrivent :

$$\begin{aligned}\tilde{\theta}_{(i)} &= x_i \\ \tilde{\theta} &= \frac{1}{n}\sum \hat{\theta}_{(-i)} = \hat{\theta}\end{aligned}$$

D'où la variance de l'estimateur de la moyenne :

$$\begin{aligned}s^2 &= \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 \\ &= \frac{1}{n(n-1)}\sum(x_i - \bar{x})^2\end{aligned}$$

Résultat fort connu !

b l'estimateur fonction de la moyenne

Nous étudierons un estimateur fonction de la moyenne : $T = g\left(\bar{X}\right)$.

Notations : $\mu = E\left(\bar{X}\right)$ γ=var$\left(\bar{X}\right)$

Nous admettrons que g est dérivable et $E(g(X)) \simeq g(E(X))$

Nous pouvons obtenir une estimation de var(T) par la méthode suivante :

$$\begin{aligned} g(x) &= g(\mu) + (x-\mu)g'(\mu) + o\left((x-\mu)^2\right) \\ E\left(\left(g\left(\bar{X}\right) - g(\mu)\right)^2\right) &\simeq E\left(\left(\bar{X}-\mu\right)^2\right)(g'(\mu))^2 \\ \text{var}\left(g\left(\bar{X}\right)\right) &= \text{var}(T) = (g'(\mu))^2 \text{var}\left(\bar{X}\right) \end{aligned}$$

Remarque :

C'est cette expression que l'on utilise pour stabiliser la variance.
Les transformations : Log, Arcsin$\sqrt{\ }$... étant obtenues en résolvant l'équation différentielle :
$K = (g'(\mu))^2 \text{var}\left(\bar{X}\right)$ avec $\text{var}\left(\bar{X}\right) = k\mu^2$; $\text{var}\left(\bar{X}\right) = p(p-1)\ldots$
K constante positive

Si nous utilisons les "pseudo-valeurs" :

$$\begin{aligned} \hat{\theta}_{(-i)} &= g\left(\frac{n\bar{x} - x_i}{n-1}\right) = g\left(\bar{x} + \frac{\bar{x} - x_i}{n-1}\right) \\ &= g(\bar{x}) + g'(\bar{x})\frac{\bar{x} - x_i}{n-1} + o\left(\left(\frac{\bar{x} - x_i}{n-1}\right)^2\right) \\ \hat{\theta}_{(.)} &= g(\bar{x}) + \sum_i o\left(\left(\frac{\bar{x} - x_i}{n-1}\right)^2\right) \end{aligned}$$

D'où :

$$\begin{aligned} \hat{\theta}_{(-i)} - \hat{\theta}_{(.)} &\simeq g'(\bar{x})\frac{\bar{x} - x_i}{n-1} \\ \frac{n-1}{n}\sum_i \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2 &= \frac{n-1}{n}(g'(\bar{x}))^2 \sum_i \frac{(\bar{x} - x_i)^2}{(n-1)^2} \\ &= (g'(\bar{x}))^2 \sum_i \frac{(\bar{x} - x_i)^2}{n(n-1)} \end{aligned}$$

Nous voyons l'analogie avec l'estimateur obtenu par la première méthode.

c l'estimateur de la variance

Considérons l'estimateur de la variance :

$$V = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Cet estimateur est non biaisé, et l'estimateur du Jackknife $\tilde{V}$ est égal à V.

La valeur prise par cet estimateur s'écrit :

$$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Les "pseudo-valeurs" s'écrivent :

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

La valeur prise par la variance de l'estimateur est :

$$s^2 = \frac{1}{n(n-1)} \sum \left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2$$

Si nous définissons le moment centré d'ordre k :

$$\begin{aligned} \mu_k &= E(X - E(X))^k \\ \text{et} \quad \hat{\mu}_k &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k \end{aligned}$$

Nous obtenons :

$$s^2 = \frac{n^2}{(n-1)(n-2)^2} \left(\hat{\mu}_4 - \hat{\mu}_2^2\right)$$

Or la vraie variance s'écrit :

$$\text{var}(V) = \frac{\mu_4 - \frac{n-3}{n-1}\mu_2^2}{n}$$

Soit A le coefficient d'aplatissement de Pearson, nous avons : $A = \frac{\mu_4}{\mu_2^2}$.

Nous pouvons réécrire les valeurs des variances :

$$\begin{aligned} s^2 &= \frac{n^2}{(n-1)(n-2)^2} \left(\hat{A} - 1\right) \hat{\mu}_2^2 \\ \text{var}(V) &= \frac{A - \frac{n-3}{n-1}}{n} \mu_2^2 \end{aligned}$$

Nous voyons ainsi que le rapport $\frac{s^2}{\text{var}(V)}$ dépend de la taille de l'échantillon et du coefficient d'aplatissement :

Le tableau suivant présente quelques rapports, pour différentes valeurs de n et A :

n \ A	1,5	1,8	3
5	1,74	2,14	2,78
10	1,20	1,36	1,56
15	1,11	1,21	1,33
20	1,07	1,15	1,23
100	1,01	1,03	1,04

Nous voyons que ce rapport diminue inversement proportionnelle à n et A.

Rappel :

Pour une loi Normale $\mathcal{N}(0,1)$, nous avons : $\mu_2 = 1$, $\mu_4 = 3$, $A = 3$.
Pour une loi Uniforme comprise entre $[a-h, a+h]$: $\mu_k = \frac{1}{k+1} h^k$, $A = \frac{9}{5}$. (1,8)

d l'estimateur de l'écart-type

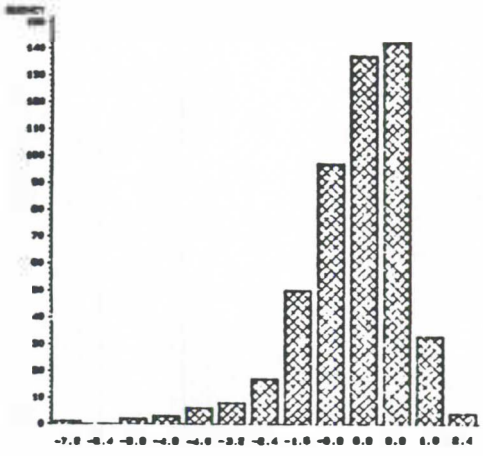
Nous reprenons les simulations effectuées au premier chapitre et, pour chacune d'entre-elles, nous calculons :

$$t = \frac{\tilde{s} - \sqrt{10}}{\sqrt{\frac{1}{n(n-1)} \sum \left(\tilde{s}_{(i)} - \tilde{s}\right)^2}}$$

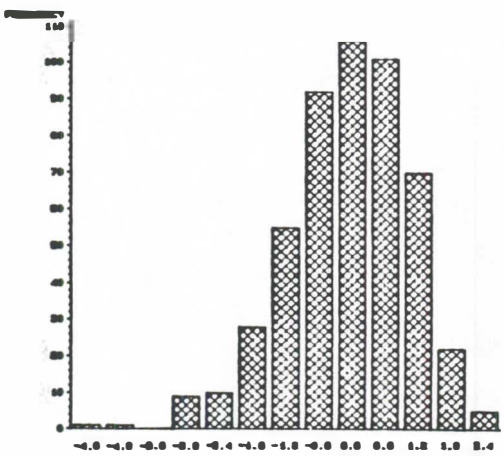
Nous obtenons ainsi 500 valeurs de t : $t_1, t_2, \ldots, t_{500}$ dont nous pouvons étudier la distribution.

En observant les histogrammes, nous constatons pour n = 16, 50 et 100 que les hypothèses formulées par Tuckey semblent réalistes (évolution vers une loi Normale).

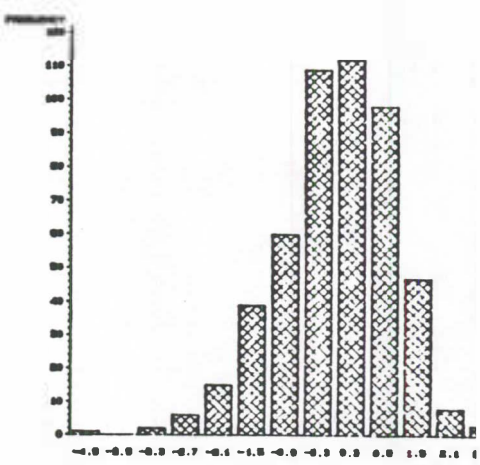
n = 16



n = 50



n = 100



e l'estimateur d'un quotient

Reprenons l'exemple présenté au chapitre précédent, concernant l'estimation de la surface boisée nationale.

La variance de l'estimateur du Jackknife est égale à :

$$\begin{aligned} s^2 &= \frac{1}{n(n-1)} \sum \left(\tilde{q}_{(i)} - \tilde{q}\right)^2 = \frac{1}{10 * 9} 1150,128 \\ &= 12,779 \end{aligned}$$

Pour valider ce résultat, nous avons tiré aléatoirement 100 échantillons de 10 départements parmi les 90. Pour chacun de ces échantillons, nous avons calculé l'estimateur du quotient et celui du Jackknife.

Les résultats sont présentés dans le tableau suivant :

Estimateur	Moyenne	Minimum	Maximum	Variance
Q	24,754	13,86	39,78	16,70
$\bar{Q}$	24,775	13,80	40,14	17,33

L'estimateur du Jackknife est légèrement plus proche de la vraie valeur que l'estimateur du quotient.
Les variances observées sont semblables à l'estimation de la variance proposée par Tukey.

f méthode de "capture-recapture"

Soit N le nombre d'individus dans une population, ce nombre est inconnu, voire inobservable.
Nous supposons qu'il existe deux listes ou structures, qui représentent chacune une partie de la population. Ces deux listes sont indépendantes.

		Seconde Liste Présent	Seconde Liste Absent	
Première Liste	Présent	N_{11}	N_{12}	$N_{1.}$
	Absent	N_{21}	–	
		$N_{.1}$		

Le nombre N_{22} est inobservable, et par conséquent la taille de la population N aussi.

Si nous supposons que ces données proviennent d'une loi multinomiale, la valeur prise par l'estimateur du maximum de vraisemblance s'écrit :

$$\hat{n} = \frac{n_{1.}n_{.1}}{n_{11}}$$

L'exemple présenté concerne l'estimation de la taille de la population agée de 14 à 64 ans des Etats-Unis [12].

La première liste représente une enquête de la population en février 1978, la seconde liste représente les impôts perçus durant l'année 1978. Ces listes seront notées respectivement "E" et "I".

Le plan d'échantillonnage de l'enquête est complexe. Les résultats proviennent de huit sous-échantillons, qui fournissent chacun des estimations de la population nationale. Les données provenant de "I" sont connues avec exactitude.

Ainsi, nous observons :

		I Présent	I Absent	
E	Présent	$\hat{n}_{11i}$	$\hat{n}_{12i}$	$\hat{n}_{1.i}$
	Absent	$\hat{n}_{21i}$	–	
		$n_{.1}$		

pour $i = 1, 2, \ldots 8$ qui correspondent aux groupes sur lesquels ont été faites les estimations. $n_{.1}$ n'est pas une estimation, mais un dénombrement.

Nous obtenons ainsi, le tableau global :

		I		
		Présent	Absent	
E	Présent	$\hat{n}_{11}$	$\hat{n}_{12}$	$\hat{n}_{1.}$
	Absent	$\hat{n}_{21}$	–	
		$n_{.1}$		

avec :

$$\hat{n}_{11} = \frac{1}{8}\sum_{i=1}^{8} \hat{n}_{11i} \quad \hat{n}_{12} = \frac{1}{8}\sum_{i=1}^{8} \hat{n}_{12i}$$
$$\hat{n}_{21} = \frac{1}{8}\sum_{i=1}^{8} \hat{n}_{21i} \quad \hat{n}_{1.} = \frac{1}{8}\sum_{i=1}^{8} \hat{n}_{1.i}$$

Les données sont présentées dans le tableau suivant :

Groupe	$\hat{n}_{11h}$	$\hat{n}_{1.h}$
1	107.285.040	133.399.520
2	105.178.160	132.553.952
3	110.718.448	139.055.744
4	103.991.496	132.390.240
5	106.818.488	131.627.520
6	106.636.928	133.095.536
7	105.338.552	133.324.528
8	103.349.328	131.061.688
Moyenne	106.164.555	133.313.591

Le nombre d'individus "présents" dans la liste "I" est : $n_{.1} = 115.090.300$

L'estimateur de la méthode de "capture-recapture" est :

$$\hat{n} = \frac{\hat{n}_{1.}n_{.1}}{\hat{n}_{11}} = \frac{(133.313.591)(115.090.300)}{106.164.555} = 144.521.881$$

Nous éliminons le $i^{ème}$ groupe, et nous calculons les nouvelles estimations :

$$\hat{n}_{(-i)} = \frac{n_{1.(-i)}n_{.1}}{\hat{n}_{11(-i)}}$$

avec:

$$\hat{n}_{11(-i)} = \frac{1}{7}\sum_{i' \neq i} \hat{n}_{11i'}$$
$$\hat{n}_{1.(-i)} = \frac{1}{7}\sum_{i' \neq i} \hat{n}_{1.i'}$$

Les "pseudo-valeurs" s'écrivent :

$$\tilde{n}_{(i)} = 8\hat{n} - 7\hat{n}_{(-i)}$$

Les résultats sont présentés dans le tableau suivant :

Groupe	$\hat{n}_{(-i)}$	$\tilde{n}_{(i)}$
1	144.726.785	143.087.553
2	144.447.797	145.040.467
3	144.518.186	144.547.744
4	144.243.095	146.473.380
5	144.910.512	141.801.461
6	144.647.594	143.641.892
7	144.359.733	145.656.914
8	144.323.897	145.907.770
Moyenne		144.519.648

L'estimateur du Jackknife est égal à :

$$\tilde{n} = \frac{1}{8}\sum_{i=1}^{8}\tilde{n}_{(i)} = 144.519.648$$

Nous pouvons calculer la variance et l'écart-type de l'estimateur du Jackknife :

$$\begin{aligned} s^2 &= \frac{1}{8*7}\sum_{i=1}^{8}\left(\tilde{n}_{(i)} - \tilde{n}\right)^2 = 3,1284 * 10^{11} \\ s &= 559.321 \end{aligned}$$

ainsi que la variance "conservatrice" :

$$s'^2 = \frac{1}{8*7}\sum_{i=1}^{8}\left(\tilde{n}_{(i)} - \hat{n}\right)^2 = 3,1284 * 10^{11}$$

Annexes

A–1 La démonstration d'Efron

Nous voulons démontrer que $E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_{n-1}$.

L'idée de base est d'écrire $T(X_1, X_2, \ldots X_n)$ comme une somme de plusieurs fonctions de une, deux, trois, ... n variables, ces variables étant indépendantes.

Nous posons :

$$\mu = E(T(X_1, X_2, \ldots X_n)) = E(T)$$

$$A_i = nE(T \mid X_i)$$
fonction uniquement de X_i

$$B_{ii'} = n^2 [E(T \mid X_i, X_{i'}) - E(T \mid X_i) - E(T \mid X_{i'}) + \mu]$$
fonction uniquement de $X_i X_{i'}$

$$N_{1,2,\ldots n} = n^n [T - E(T \mid X_1 \ldots X_n) - \cdots - (-1)^n \mu]$$

Nous avons : $E(A_i) = E(B_{ii'}) = \cdots E(N_{1,2,\ldots n}) = 0$.
Les variables $A_i, B_{ii'}, \ldots$ sont non corrélées.

Enfin :

$$T_n = \mu + \frac{1}{n}\sum_i A_i + \frac{1}{n^2}\sum_{ii'} B_{ii'} + \cdots \frac{1}{n^n} N_{1,2,\ldots n}$$

Notons :

$$\sigma_A^2 = \text{var}(A_i) \quad \sigma_B^2 = \text{var}(B_{ii'}) \quad \text{etc}\ldots$$

et considérons la décomposition de T_n par rapport à T_{n-1}.

a)

Nous pouvons considérer cette décomposition par rapport à $T_{(-i)}$ qui est construit à partir d'un $(n-1)$-échantillon.

$$T_{(-i)} = \mu^i + \frac{1}{n-1}\sum_{j \neq i} A_j^i + \frac{1}{(n-1)^2} \sum_{\substack{j<j' \\ j\neq i \text{ et } j'\neq i}} B_{jj'}^i + \cdots$$

Remarque :

$\mu^i = E\left(T_{(-i)}\right)$ est en fait l'espérance de l'estimateur T_{n-1} construit sur un $(n-1)$ - échantillon
$A_j^i = E\left(T\left(X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots X_n\right) \mid X_j\right) = E\left(T_{n-1} \mid X_j\right)$
etc ...
Ce qui signifie que $\mu^i, A_j^i, \ldots$ sont indépendants de i et peuvent être notés μ, A_j (ces variables ne sont pas indépendantes de n).

Remarque :

$$\text{var}\left(T_{(-i)}\right) = \gamma_{n-1} = \frac{\sigma_A^2}{n-1} + C_{n-2}^1 \frac{\sigma_B^2}{2(n-1)^3} + C_{n-2}^2 \frac{\sigma_C^2}{3(n-1)^5} + \cdots$$

b)

Si nous remarquons que :

$$\sum_i \left(T_{(-i)} - T_{(.)}\right)^2 = \frac{1}{n} \sum_{i<i'} \left(T_{(-i)} - T_{(-i')}\right)^2$$

Nous voyons :

$$E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) = \frac{1}{n} E\left(\sum_{i<i'} \left(T_{(-i)} - T_{(-i')}\right)^2\right) = \frac{n-1}{2} \text{var}\left(T_{(-i)} - T_{(-i')}\right)$$

Nous allons chercher à exprimer $\left(T_{(-i)} - T_{(-i')}\right)$ en fonction des $A, B \ldots$ et à calculer var $\left(T_{(-i)} - T_{(-i')}\right)$ en fonction des $\sigma_A^2, \sigma_B^2, \ldots$ ce qui permettra de comparer $E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right)$ à γ_{n-1}.

c)

Etudions le comportement des termes A_j dans la différence $\left(T_{(-i)} - T_{(-i')}\right)$.

Lorsque nous effectuons la différence $\left(T_{(-i)} - T_{(-i')}\right)$ il ne reste que les $A_{i'} - A_i$
De même pour les termes en B, il ne reste que $B_{ji} - B_{ji'}$ etc ...
D'où :

$$\begin{aligned}
\left(T_{(-i)} - T_{(-i')}\right) &= \frac{1}{n-1}\left(A_{i'} - A_i\right) + \frac{1}{(n-1)^2} \sum \left(B_{ji'} - B_{ji}\right) + \cdots \\
\text{var}\left(T_{(-i)} - T_{(.)}\right) &= 2\left(\frac{1}{(n-1)^2}\sigma_A^2 + \frac{1}{(n-1)} C_{n-2}^1 \sigma_B^2 + \frac{1}{(n-1)} C_{n-2}^2 \sigma_C^2 + \cdots\right) \\
\text{et} & \\
E\left(\left(T_{(-i)} - T_{(.)}\right)^2\right) &= \frac{n-1}{2} \text{var}\left(T_{(-i)} - T_{(-i')}\right) \\
&= \frac{\sigma_A^2}{n-1} + \frac{C_{n-2}^1}{(n-1)^3}\sigma_B^2 + \frac{C_{n-2}^2}{(n-1)^5}\sigma_C^2 + \cdots
\end{aligned}$$

d)

Si nous comparons maintenant terme à terme :

$$E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \quad \text{et} \quad \gamma_{n-1}$$

Nous voyons que :

$$E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_{n-1}$$

et la différence est égale à :

$$\frac{1}{2}\frac{C_{n-2}^1}{(n-1)^3}\sigma_B^2 + \frac{2}{3}\frac{C_{n-2}^2}{(n-1)^4}\sigma_C^2 + \cdots$$

Remarque :

Les termes $\mu, \sigma_A^2, \sigma_B^2, \ldots$ sont des fonctions de n (hélas !).

Bibliographie

[1] Ministère de l'agriculture. Statistiques forestières en 1983. *Service Central des Enquêtes et Etudes Statistiques*, 30, Avril 1985.

[2] J. Durbin. A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46:477–480, 1959.

[3] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[4] H.L. Gray and W.R. Schucany. *The Generalized Jackknife Statistic.* Marcel Dekker, New-York, 1972.

[5] W. Hoeffding. A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.*, 19:293–325, 1948.

[6] M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin and Company limited, London, third edition, 1969.

[7] Rupert G. Miller. The Jackknife - a review. *Biometrika*, 61:1–15, 1974.

[8] F. Mosteller and J.W. Tukey. *Data Analysis and Regression.* Addison Wesley, 1977.

[9] M. Quenouille. Approximate tests of correlation in time series. *J. Roy. Statist. Soc. B*, 11:68–84, 1949.

[10] M. Quenouille. Notes on bias in estimation. *Biometrika*, 43:353–360, 1956.

[11] J.W. Tukey. Bias confidence in not quite large samples. *Ann. Math. Statist.*, 29, 1958. 614.

[12] K. M. Wolter. *Introduction to Variance Estimation.* Springer-Verlag, New-York, 1985.